Unifying Vision-and-Language Tasks via Text Generation

Year: 2021 | Citations: 602 | Authors: Jaemin Cho, Jie Lei, Hao Tan, Mohit Bansal

---

Abstract

Existing methods for vision-and-language learning typically require designing task-specific architectures and objectives for each task. For example, a multi-label answer classifier for visual question answering, a region scorer for referring expression comprehension, and a language decoder for image captioning, etc. To alleviate these hassles, in this work, we propose a unified framework that learns different tasks in a single architecture with the same language modeling objective, i.e., multimodal conditional text generation, where our models learn to generate labels in text based on the visual and textual inputs. On 7 popular vision-and-language benchmarks, including visual question answering, referring expression comprehension, visual commonsense reasoning, most of which have been previously modeled as discriminative tasks, our generative approach (with a single unified architecture) reaches comparable performance to recent task-specific state-of-the-art vision-and-language models. Moreover, our generative approach shows better generalization ability on questions that have rare answers. Also, we show that our framework allows multi-task learning in a single architecture with a single set of parameters, achieving similar performance to separately optimized single-task models. Our code is publicly available at: https://github.com/j-min/VL-T5